TN-02: Statistical Comment on the Study Concept/Prototype

Joan F. Hilton, Sc.D.
Associate Professor of Biostatistics
University of California San Francisco

March 16, 2005

Preliminary Questions:

1. How long should the treatment and follow-up period in trials be?
2. *Can larger numbers of treatments be compared simultaneously with aggressive curtailment for futility?*
3. Can short-term trials be grafted onto long-term trials - roll Phase II patients into Phase III studies?
4. *Should trials be powered for large effect sizes or for more moderate effect sizes with monitoring guidelines that could terminate early for an unexpected large effect?*
5. *Should trials be powered to rule out adverse effects?*

$$\left.\begin{array}{l}\text{Duration of a trial (Q1, Q3)}\\ \text{Number of treatments (Q2)}\\ \text{Number of patients (Q4, Q5)}\end{array}\right\}\rightarrow\begin{array}{c}\text{All depend on}\\ \text{the outcome measure!}\end{array}$$

How can we amplify the value of the outcome measure in TN-02*?*

Idea 1: Use composite endpoints.

Example: The Women's Health Initiative prevention trial.

Idea 2: Use response-adaptive allocation procedure.

Example: AML trial comparing CR rates of 2 Experimental treatments with a Control.

Idea 1: Use composite endpoints.

The Women's Health Initiative (WHI) Argument:

Trials addressing treatment of established disease → outcome should be univariate.

Trials addressing prevention / early disease → outcome should be multivariate.

Rationale:

Treatment trials	**Prevention trials**
• Given established disease, intervention should alleviate direct consequences	• Given generally healthy, early disease-related morbidity & mortality are rare; intervention should maintain health
• Intervention benefits should outweigh risks	• Intervention risks may outweigh benefits
• Intervention effects are typically known; Trade early benefits for late risks	• Intervention effects may be unknown; Time courses of benefits and risks may differ
• The study is typically replicated	• The study is too large to be replicated
→ **Specify a single primary outcome** Estimate primary effect	→ **Consider a collection of endpoints** Estimate broad range of effects

Other WHI and *TN-02* Analogies:

Subjects are randomized to more than one intervention component.

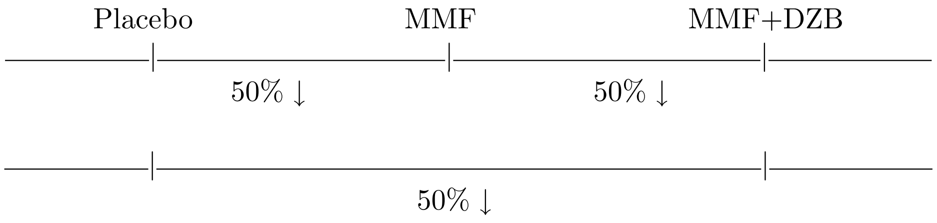
The intervention components may interact.

Intervention components and their outcomes

WHI	**TN-02**
Low-fat diet Vs. Usual diet	**MMF Vs. Placebo**
↓ CHD	Affects β cells via B-lymphocytes
↓ Colorectal cancer	Prevents islet allograft rejection
↓ Breast cancer	SAEs
Hormone replacement Vs. Placebo	**DZB Vs. Placebo**
↓ CHD	Affects β cells via T-lymphocytes
↑ Breast cancer	Halts graft-v-host disease
↓ Hip fractures	
↓ Endometrial cancer	

Differences in β-cell retention:

Protocol specifies 3 Superiority tests:

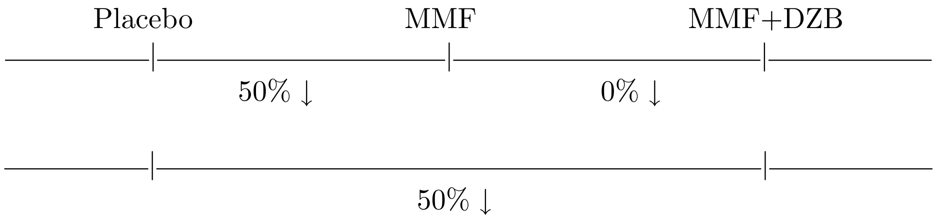


Placebo MMF MMF+DZB

50% ↓ 50% ↓

50% ↓

Alternatively, conduct 2 Superiority tests and 1 Noninferiority test:



Placebo MMF MMF+DZB

50% ↓ 0% ↓

50% ↓

- For MMF versus MMF+DZB, study DZB effect.
- What outcome variable focuses on DZB effect?

WHI Recommendations:

1. Measure benefit in more than one way.

- Disease-specific outcome: *β cell retention*
- Global health outcome: *A variety of surrogate markers?*
  - Unweighted combination of treatment outcomes
  - Weighted combination of treatment outcomes; weights
    - *prediction of T1DM*
    - strength of evidence for prediction

Aside 1: Peter O'Brien's Global Test.

1. Rank each outcome across all subjects in terms of efficacy: $R_{j,1}, \ldots, R_{j,n_E+n_C}$, for $j = 1., \ldots, J$.
2. Replace the raw data with the ranks.
3. For each patient, sum the ranks across outcomes: $S_i = R_{1,i} + R_{2,i} + ... + R_{J,i}$, for $i = 1, ..., n_E + n_C$.
4. Test H_0 using these summary data.

Advantage: Reduce a J-dim to a 1-dim outcome per subject; independent observations.

Disadvantage: To interpret, supplement with sub-groups of tests or individual tests.

Note: A large P-value might suggest that some component measures are not sensitive to efficacy, whereas the sub-group analysis might identify those that are and should be studied further.

Aside 2: Latent Variable Modelling.

Advantage: Component measures are adjusted for one another but retain interpretability.

2. Measure risk in more than one way:

- Intervention-targetted outcomes
- Global adverse events

3. Formal stopping rules: Stop if *either* benefit *or* harm.

- α-levels may vary. Efficacy examples:
  - Disease-specific outcome: $\alpha = 0.05$
  - Global health outcome: $\alpha = 0.20$ ("supportive" evidence)
- Boundaries for benefit and harm needn't be symmetric

4. Monitor frequently.

Planned follow-up in the WHI Trial was 8.5 years.

After 5.2 years' mean follow-up, stopped at the 10^{th} analysis:

- The boundary for adverse events was crossed.
- The global index supported the conclusion of harm.

Idea 2: Use response-adaptive allocation procedures.

Example: AML trial comparing CR rates of 2 Experimental treatments with a Control.

Setting: Patients evaluated very soon after randomization for a binary outcome, Complete Response (CR).

Initial allocation probabilities:

Fixed allocation to C, 33%;

Adaptive allocation to E_1 versus E_2, based on relative CR rate (π).

C = idarubicin and cytarabine; E_1 = troxacitabine and idarubicin;
E_2 = troxacitabine and cytarabine

Pt	Pr$\{C\}$	n_C	π_C	Pr$\{E_1\}$	n_{E_1}	π_{E_1}	Pr$\{E_2\}$	n_{E_2}	π_{E_2}
1	0.33	0	.	0.33	1	0	0.33	0	.
2	0.33	1	1.000	0.32	1	.	0.34	0	.
3	0.33	1	.	0.32	2	0	0.35	0	.
4	0.33	2	0.500	0.30	2	.	0.37	0	.
5	0.33	3	0.333	0.28	2	.	0.38	0	.
6	0.33	4	0.500	0.28	2	.	0.39	0	.
7	0.33	5	0.400	0.27	2	.	0.39	0	.
8	0.33	5	.	0.23	3	0	0.44	0	.
9	0.33	5	.	0.20	4	0	0.47	0	.
10	0.33	5	.	0.24	4	.	0.43	1	1.000
11	0.33	5	.	0.17	4	.	0.50	2	0.500
12	0.33	5	.	0.17	4	.	0.50	3	0.333
13	0.33	5	.	0.20	4	.	0.47	4	0.250
14	0.33	5	.	0.10	5	0	0.57	4	.
15	0.33	5	.	0.10	5	.	0.57	5	0.400
16	0.33	6	0.333	0.11	5	.	0.56	5	.
17	0.33	6	.	0.11	5	.	0.56	6	0.500
18	0.33	6	.	0.11	5	.	0.55	7	0.429
19	0.33	6	.	0.13	5	.	0.54	8	0.375
20	0.33	7	0.429	0.14	5	.	0.53	8	.
21	0.33	8	0.500	0.18	5	.	0.49	8	.
22	0.33	9	0.556	0.21	5	.	0.46	8	.
23	0.33	10	0.600	0.09	5	.	0.58	8	.
24	0.33	11	0.636	0.07	5	.	0.59	8	.

Late allocation probabilities:

After patient 24 responded, drop Arm E_1 (lack of efficacy).

Adaptive allocation to C versus E_2, based on relative CR rate (π).

Potential problem:

No guarantee that groups are balanced by prognostic covariates

Patients in E_1 could be at highest risk of failure.

Solution:

Use group-sequential methods instead: retain advantages of randomization, but monitor results and drop arms as warranted

C = idarubicin and cytarabine; E_1 = troxacitabine and idarubicin;
E_2 = troxacitabine and cytarabine

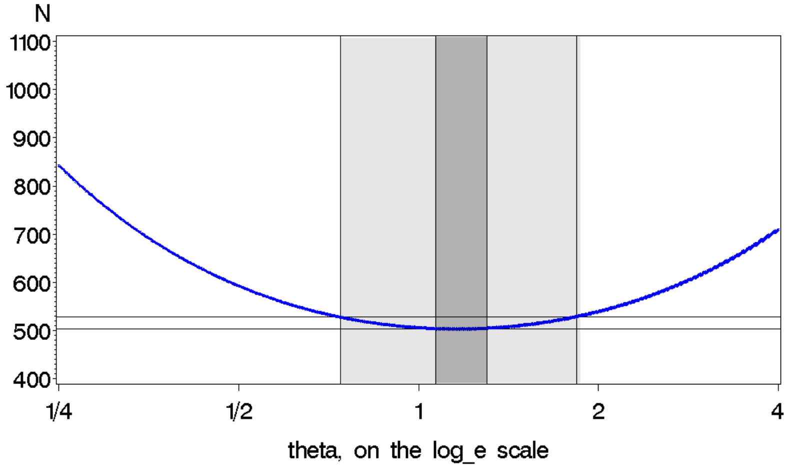
Pt	$\Pr\{C\}$	n_C	π_C	$\Pr\{E_1\}$	n_{E_1}	π_{E_1}	$\Pr\{E_2\}$	n_{E_2}	π_{E_2}
25	0.87	12	0.583	–	5	.	0.13	8	.
26	0.87	12	.	–	5	.	0.13	9	0.333
27	0.96	12	.	–	5	.	0.04	10	0.300
28	0.96	13	0.615	–	5	.	0.04	10	.
29	0.96	14	0.571	–	5	.	0.04	10	.
30	0.96	15	0.600	–	5	.	0.04	10	.
31	0.96	16	0.562	–	5	.	0.04	10	.
32	0.96	16	.	–	5	.	0.04	11	0.273
33	0.96	17	0.529	–	5	.	0.04	11	.
34	0.96	18	0.556	–	5	.	0.04	11	.
		18	0.556		5	0.0		11	0.273

Potential problem:

What effect does the sample-size imbalance have on power level?

Figure 1: Example. For test statistic based on $\hat{\Delta}$, with $\{\pi_C, \Delta\} = \{0.03, 0.05\}$ and $\alpha = 0.05$, $\beta = 0.20$, the overall sample size, $N = n_E + n_C$

- is at its minimum (dark shading) when $\theta = n_E/n_C \in (0.73, 1.35)$
- is within 5% of its minimum (light shading) when $\theta = n_E/n_C \in (0.53, 1.54)$.



Giles Trial – final sample size ratios:

- $\theta_1 = n_{E_1}/n_C = 0.278 \Rightarrow$ low power
- $\theta_2 = n_{E_2}/n_C = 0.611 \Rightarrow$ probably little power loss; specific trial parameters $\{\pi_C, \Delta\} = \{0.55, 0.30\}$ should be evaluated

Point: For some nonnegligible imbalance in sample sizes across groups, there is little power loss. As imbalance increases, power loss becomes more dramatic.

Analogies:

TN-02 Setting: Patients evaluated periodically for a continuous outcome, β-cell retention, which may continue to change over a long period.

Research challenge: Could an adaptive randomization procedure be based on a multivariate measure?

Throughout the trial: Can allocation to the most effective treatment be based on evolving outcome data?

Summary:

Two techniques to reduce sample size and/or study duration were discussed.

Remaining challenges:

- How to define multivariate benefit & risk outcomes and/or univariate global measure.
  - WHI 8.5-year trial: Used incidences of several clinical events.
  - *TN-02* shorter trial: Use a collection of auxiliary endpoints? Seek consistent, biogically plausible evidence.
- How to use these outcomes in a response-adaptive randomization procedure and/or Group-sequential procedure: Frequent monitoring of multivariate outcome should distinguish among groups.

References:

Idea 1:

Freedman L, Anderson G, Kipnis V, et al. Approaches to monitoring the results of long-term disease prevention trials: examples from the Women's Health Initiative. *Control Clin Trials*. 1996, 17:509-25.

Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*. 2002, 288:321-33.

Idea 2:

Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science*. 2004, 19:175-187.

Giles FJ, Kantarjian HM, Cortes JE, et al. Adaptive randomized study of idarubicin and cytarabine (C) versus troxacitabine and idarubicin (E_1) versus troxacitabine and cytarabine (E_2) in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. *J Clin Oncol*. 2003, 21:1722-7.